DOCUMENT RESUME

ED 135 852 TM 006 083

AUTHOR Millman, Jason
TITLE Creating Domain-Referenced Tests by Computer.
PUB DATE [Apr 77]
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Computer Programs; *Criterion Referenced Tests; *Item Banks; Mastery Tests; *Test Construction
IDENTIFIERS *Domain Referenced Tests

ABSTRACT

A unique system is described for creating tests by computer. It is unique because, instead of storing items in the computer, item algorithms similar to Hively's notion of item forms are banked. Every item, and thus every test, represents a sample from domains consisting of thousands of items. The paper contains a discussion of the special practical applications of such tests, a description of the easy-to-learn user language in which item algorithms are written, and the results of using the tests in a college course taught by the mastery learning instructional strategy. (Author/RC)

Creating Domain-Referenced Tests by Computer[1]

Jason Millman
Cornell University

The standard way computers are used to generate tests requires the user to store items in the computer exactly as they are to appear on the test. Typically, the test constructor stores or "banks" several hundred test items together with some cataloging information such as the course topic measured by the item. A test is constructed by sampling questions from the population of items banked in the machine.

In the system discussed in this paper, item programs rather than the items themselves are stored. An item program is a computer program written in a specially constructed, easy-to-learn language. Each item program, in turn, is capable of producing multiple items, that is, different versions of questions about a given knowledge or skill area. Thus, when a collection of item programs is stored in the computer, the potential for creating vast numbers of items is present. Every test composed of items generated from the item programs will be unique because it is unlikely that specific test questions will be repeated.

Specific applications are presented first. Next, considerations bearing on the use of the system will be discussed. The steps the educator follows in using the system and the author's experience with it will then be described. Finally, brief mention will be made of needed future development.

Specific Applications

Recall that the computerized test construction system is capable of

[1]A paper presented at the annual meetings of the American Educational Research Association and the National Council for Measurement in Education, New York City, April 1977.

producing unique tests. Although the items on each test will be created from the same group of item programs, their specific content will differ. It is this feature that makes the computerized system particularly attractive.

Repeated testing. Mastery learning and the personalized system of instruction are but two modes of teaching in which students are permitted multiple tries at demonstrating competence with respect to unit or course goals. The computer generated tests are especially convenient in this situation or for any occasion when repeated testing of the same individuals is permitted. The tests taken during the second, third, or later trials will be randomly parallel to, but different from the initially administered test. Thus, the educator has available a nearly unlimited number of different forms of each test.

In some certification, licensing, or classroom testing programs, testing is "by demand," which means that it is the client who determines when testing will take place. In this type of situation, test security is impossible. It is very important to have available a large number of test forms.

Practice materials. No classroom is ever truly homogeneous with respect to student achievement. Some students need more practice, others less. Yet the construction of practice materials, important as they are for insuring learning, is extremely time consuming. The computerized system can also be thought of as a near endless source of practice questions. Figure 1 is a reproduction of one such practice sheet. In this example, only a single item program was used.

Control of cheating. Even when an entire class is administered an examination in a single setting, using the computerized system is still advantageous. Cheating, more of a problem with objective-type examinations than we may care to admit, is eliminated since every person being tested has a different examination. The random process of selection of content in even the most simple of the item programs guarantees the futility of looking at a neighbor's answer.

Make-up tests. Exceptions can throw schedules out of kilter. In testing, the need for make-up examinations can be particularly time consuming and annoying. With computer generated tests, the unused tests will be ideal as make-up examinations.

Tailored tests. Construction of specially tailored tests is another potential application. Different instructions for producing tests can be used so that a student's test can be geared to his or her culture, reading level, interests, and so on.

Considerations

Domain-referenced tests. The practice items displayed in figure 1 constitute a domain-referenced test (Millman, 1974). The item program is an explicit definition of the population of items referenced by the test. As will be mentioned later, item programs use as guides item forms (Hively, 1962), amplified objectives (Popham, 1975), or mapping sentences (Guttman, 1969).

The importance of domain-referenced tests and the concommitant clarity with which item populations are described have been stated by many (e.g., Millman, 1974; Popham, 1974). Being clear about what students can or cannot do has been seen as especially valuable by those in charge of instruction in our schools. Virtually none of the commercially available criterion-referenced tests presently provides this clarity. Computer-generated, domain-referenced tests require it.

Equivalence of forms. Two tests generated from the same item program(s) may be considered randomly parallel but, in general, will not be equal in difficulty. This inequality reduces the usefulness of such tests for comparing examinees with each other. For the situation in which an individual's ability is being estimated (e.g., with respect to a standard), no bias or reduction in accuracy occurs by having different individuals respond to different sets of items. In fact, to estimate the mean ability level of a group, such tests are

an advantage (Lord, 1976).

Costs. The cost of the system is difficult to assess. Not only are there different types of cost but also there is no reasonable alternative for acquiring the same benefits.

One source of costs is for the computer itself. More and more schools are purchasing intermediate-size computers, and if a school already has purchased such equipment, no additional purchases may be necessary. A second source of costs are variable computer costs that include maintenence, time rental, and paper. Again, for these modest size computers, the only monetary cost for the user would be for paper. Finally, a developmental cost is associated with creating the item programs. The time needed to construct these programs would be greatly reduced if an item pool or set of objectives already existed. The entire purpose of the special language is to facilitate the development of these programs. The developmental costs are mostly "fixed" costs and thus amortizable over a number of years. Revisions to existing programs can be easily made.

Against these charges, one must consider not only the educational benefits, but other advantages that perhaps can be converted more directly into dollar savings. Once the system is in process, the testing operation can be routinized using lower-salaried personnel. Also one should take into account the time instructors would otherwise take to write or assemble their tests and the time and expense involved in typing and reproducing the tests (and make-up examinations). Although the present system does not include automatic scoring and recording procedures, these could be added and thus could result in additional savings. Emerson (1974) provides additional comments on costs associated with computer-generated tests.

Appropriateness for all subject areas. Item programs can be constructed for skill and knowledge areas such as mathematics, the sciences, grammar, and so on. Although more difficult and frequently not feasible, item programs can

be created to measure more general capabilities. For example, the items in figure 2 generated from a single program ask the student to evaluate and categorize the response of a parent to a child who has just lost someone or something of value.

Steps in Test Construction

As in any test construction effort, the beginning point is the identification of the knowledges or skills to be tested. This identification can be in the form of instructional objectives or in the form of test items that are already available to the teacher. Figure 3 contains one such test item.

The item shown in figure 3 could be stored in the computer by preparing the item program listed in figure 4. The line numbers are arbitrary except that they indicate the order the steps in the program should be executed. The statements in lines 10 and 20 designate content that will be presented to the student as part of the question. Depending upon the printing format for the test, the answer given in line 30 might not appear on the test page.

The item program displayed in figure 4 is a special case; it is capable of generating only a single version of the test item, namely the version shown in figure 3. By using such item programs, the test generation system can be used in the standard way, namely to bank individual items.

To take advantage of the system's full capability, the test constructor can compose variations of the single item. Some variations of the illustrative items have been schematized in figure 5. Such a representation of the versions in which an item can be presented is called an item form.

As indicated at point A in the item form, in figure 5, the test items could begin with "Season affects" or with "Specific nutrition deficiencies affect". Any one of four different animal pairs could be used in the item (see points B and B'). The item form also indicates that "True" and "False" should

be listed as test options, and that the correct answer depends upon the choices made at points A and B.

Thanks to a specially constructed language, the conversion of an item form to an item program is straightforward. Figure 6 contains the item program corresponding to the item form shown in figure 5. The computer statement, FROM, instructs the machine to pick from among the following options in accordance with the directions given in a CHOOSE statement. For example, line 40 directs the computer to form a test item by choosing at random either the content of line 20 or the content of line 30. Line 180, on the otherhand, instructs the computer to pick the same choice (first, second, third, or fourth option) as was chosen when line 110 was executed. In that way, "bulls" and "cows", "boars" and "sows", and so on will be paired. The LIST 2 statement in line 190 signals the computer that the next two statements should be listed in a multiple-choice format in the order given.

Eight different items can be produced from this program because one of two choices at point A in the item form can be combined with one of four choices at point B. Four of these eight versions are shown in figure 7. With more elaborate item forms, very large numbers of versions can be produced from a single program.

Millman and Outlaw (in press) present much more information about the computer language the test constructor can use to convert item forms to item programs. The language has been designed to facilitate this conversion for a wide variety of item types and subject matter. The purpose of the language has been to lighten substantially the work involved in this step of the test development process.

After the item program has been written it can be inserted into the computer many different ways. One method involves the test constructor or secretary sitting at a console and typing in the material much as one would

do at a typewriter. At present, no interaction takes place between the computer and a student, although the system could be extended to make this possible.

Once the item programs are in the computer, the next step is to have the computer generate multiple versions of each item similar to those shown in figure 7. A display of these items will reveal possible errors in the program, errors which can be easily corrected by retyping the faulty lines.

As a final step, the tests themselves are generated. By answering a series of questions posed by the computer, the test user specifies exactly which item programs are to be sampled (the test content) and how the test is to appear on the page (the test format).

The test construction steps are summarized in figure 8.

Creating Statistics Tests

During the spring 1976 school term, the author constructed 132 item programs that formed the basis for seven mastery tests. Each mastery test covered about two weeks work. Items 6-12 on the first mastery test dealt with the summation operator. Slightly reduced reproductions of this portion of two different tests are shown in figures 9 and 10 to illustrate the variation present between randomly parallel tests.

Students did very well on the mastery tests. Although they were permitted six attempts, each test was administered only 1.7 times on the average. Consequently, the author's hopes of providing evidence on the effect of mastery testing on learning were not realized. That is, since most students took any given test only once, it was not possible to assess adequately the effect of repeated testing on subsequent learning.

Nevertheless, it may be instructive to compare the group's performance on a comprehensive final with the performance of the previous year's class on the same final. The previous class did not have experience with the mastery tests, although some item types found on the mastery tests had been used as practice

exercises the year before.

On each of the two forms of the comprehensive final, the two classes performed essentially the same (within 2% of each other). As indicated in the tables below, however, differences in percent correct scores between the classes (i.e., years) resulted on subgroups of items identified either as (a) similar to items appearing in the mastery test or (b) containing content not emphasized on the mastery test.

	Items Similar to Mastery Tests			Items on Content Not Emphasized		
	1975	1976	t	1975	1976	t
Form 1	80%	90%	-1.27	85%	68%	1.48
Form 2	75%	86%	-1.52	73%	65%	.68

The data shown above are consistent with the view that the mastery tests forcused the learners' attention on selective subject matter. If true, instructors would be well advised to make the coverage of their mastery tests span the range of skills and knowledges for which they wish their students responsible. Further discussion of the study may be found in Millman (1976).

Future Developments

Additional work on the system is underway or planned. Needed expansion and modification of the language itself will become evident as different users attempt to implement the system. The system could be developed to make it compatible with a variety of computer installations, to allow greater flexibility in selecting item programs and printing test questions, and to integrate it with possible instructional and instructional management systems.

Theoretical work on and guidelines for specifying domains is critically needed. Millman (1977) discussed some of the difficulties in identifying empirically the most important domain defining task variables.

PRACTICE ITEMS ON REFERENCE SOURCES

NAME:_____

1. TO FIND THE PAGES IN A BOOK A STORY IS, THE BEST PLACE TO LOOK IS:

A. TABLE OF CONTENTS
B. ENCYCLOPEDIA
C. CARD CATALOG

2. TO FIND HOW TO PRONOUNCE FOOTNOTE, THE BEST PLACE TO LOOK IS:

A. DICTIONARY
B. TABLE OF CONTENTS
C. INDEX

3. TO FIND OTHER WORDS THAT MEAN THE SAME THING AS CONCAVE, THE BEST PLACE TO LOOK IS:

A. THESAURUS
B. TABLE OF CONTENTS
C. INDEX

4. TO FIND THE DEFINITION OF PICA, THE BEST PLACE TO LOOK IS:

A. TABLE OF CONTENTS
B. ENCYCLOPEDIA
C. GLOSSARY

5. TO FIND BOOKS ABOUT PLANTING A GARDEN, THE BEST PLACE TO LOOK IS:

A. INDEX
B. CARD CATALOG
C. DICTIONARY

Figure 1. The test generation system can be used to provide practice materials. Answers can be placed directly on the materials or printed separately.

Mother to daughter whose dog, Sneaker, has just died: YOU CAN'T FEEL THAT BADLY.

A. moralizing
B. denying experience
C. belittling
D. condescension

ANSWER: B

Son: Cathy has broken up with me.

Mother: DO YOU FEEL VERY BADLY ABOUT THAT?

A. recognizing feeling
B. taking responsibility for the feeling
C. being receptive
D. describing results

ANSWER: C

Mother to daughter whose best friend, Shirley, is moving away: IT ISN'T THAT TERRIBLE.

A. indirect attack
B. sarcasm
C. belittling
D. denying experience

ANSWER: D

Mother to son whose dog, Hecate, has just died: SO WHAT ELSE IS NEW?

A. contradictory messages
B. condescension
C. sarcasm
D. indirect attack

ANSWER: C

Figure 2. Computer generated test items referencing a higher level cognitive skill. Student is asked to identify the parent response according to Haim Ginott.

Season affects the quality of semen in bulls as well as cyclic behavior in cows. (True or False)

Figure 3. A single test item.

```
10 QUESTION CONTENT "Season affects the quality of semen in bulls "
20 QUESTION CONTENT "as well as cyclic behavior in cows.
   (True or False)"
30 ANSWER CONTENT "False"
```

Figure 4. An item program correspondng to the test item shown in figure 3.

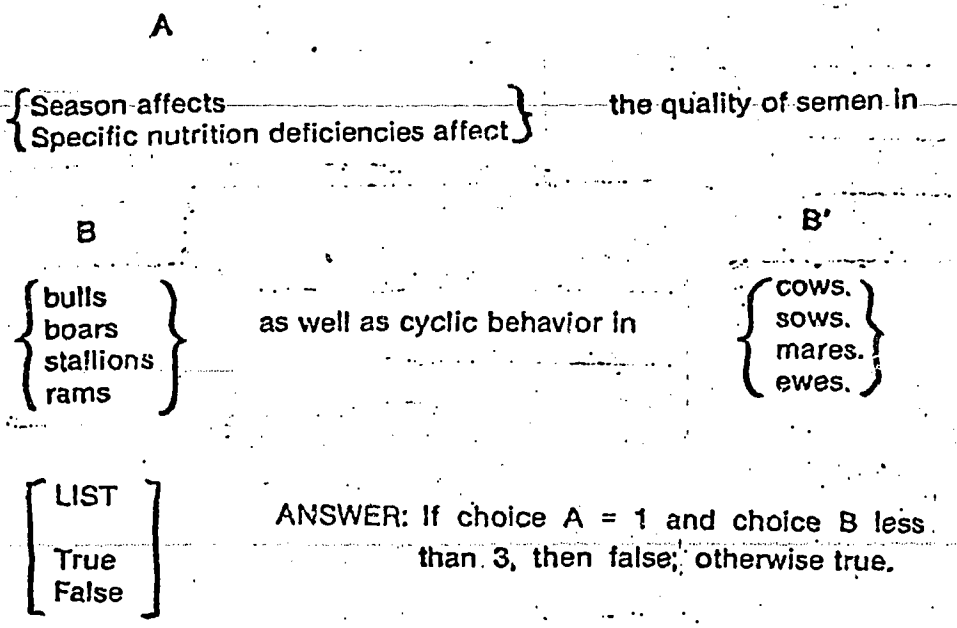

Figure 5. An item form designed to produce variations of the test item shown in figure 3.

```
 10 FROM
 20 "Season affects "
 30 "Specific nutrition deficiencies affect "
 40 CHOOSE AT RANDOM
 50 "the quality of semen in "
 60 FROM
 70 "bulls"
 80 "boars"
 90 "stallions"
100 "rams"
110 CHOOSE AT RANDOM
120 " as well as cyclic behavior in"
130 FROM
140 "cows."
150 "sows."
160 "mares."
170 "ewes."
180 CHOOSE CHOICE (110)
190 LIST 2
200 "True"
210 "False"
220 IF CHOICE (40) = 1 AND CHOICE (110) < 3 THEN ANSWER CON
    TENT "B. False"
230 ELSE ANSWER CONTENT "A. True"
```

Figure 6. An item program corresponding to the item form shown in figure 5.

Season affects the quality of semen in boars as well as cyclic behavior in sows.

A. True
B. False

ANSWER: B. False

Specific nutrition deficiencies affect the quality of semen in stallions as well as cycylic behavior in mares.

A. True
B. False

ANSWER: A. True

Specific nutrition deficiencies affect the quality of semen in rams as well as cyclic behavior in ewes.

A. True
B. False

ANSWER: A. True

Season affects the quality of semen in bulls as well as cyclic behavior in cows.

A. True
B. False

ANSWER: B. False

Figure 7. Four of the possible items that could be generated from the item program listed in figure 6.

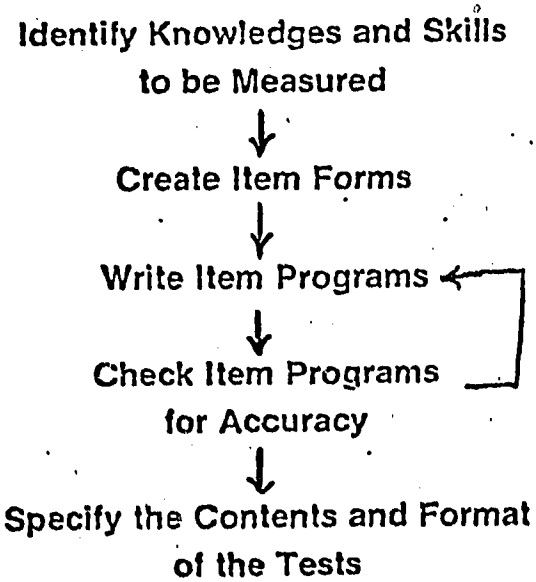

Figure 8. Steps in the computer generated, test construction process.

ID 145
PAGE 2

NAME: _____

6. 3, 2, 1 ——— 6. Assume the convention where i, j, and k are the indices for the variables designated by the rows, columns, and observations within cells. For the data below, what are the numerical values of i, j, and k, for $X_{i,j,k} = 167$

	1	2	3
1	55 67	30 68	89 47
2	58 58	40 30	36 72
3	56 86	16 71	72 97

7. 49 ——— 7. For the data shown below, what is the numerical value of: $(\Sigma Y)^2$?

X	Y
0	3
-1	3
3	1

8. A. True
1/2 point ——— 8. $\sum_i \sum_j X_{ij}^2 = \sum_i \sum_j (X_{ij})^2$

A. True
B. False

9. ΣX^2 or $\Sigma(X)^2$ ——— 9. Rewrite the arithmetical expression using summation notation. In your answer, do not show any of the numbers displayed below.

	SCORES				TOTAL
X:	8	5	7	12	32
Y:	9	11	10	3	33

Arithmetical expression: $64 + 25 + 49 + 144$

10. 114 ——— 10. Assume i is the row index. For the data shown below, what is the numerical value of:

$\Sigma\Sigma X_{ij}^2$?

5	3	5
1	2	0
4	5	3

11. 70
part cr. for
170 ——— 11. Evaluate $\sum_{i=1}^{10} (6W - 5)$ given $\sum_{i=1}^{10} W = 20$

12. A. True
1/2 point ——— 12. $\Sigma 4(b+7U) = 4bN + 28\Sigma U$ (Where N is the number of elements summed)

A. True
B. False

Figure 9. A reduced reproduction of a portion of a mastery test on the summation operator.

ID 146
PAGE 2

Educational Statistics TEST 1 (Topic A)

NAME: _____

$\Sigma(Y-X)$

6. X AND Y represent scores on a mathematics test taken before and after instruction respectively. The gain in mathematics achievement for any individual is indicated by Y - X. Write, using summation notation, X, and Y, the sum of the gain scores.

144

7. Assume i is the row index. For the data shown below, what is the numerical value of:

$$(\Sigma\Sigma X_{ij})^2_{ij} ?$$

3	4
3	0
1	1

35496

8. For the data shown below, what is the numerical value of: $\Sigma U^2 \Sigma W^2$?

	SCORES				TOTAL
U:	10	11	6	7	34
W:	1	5	3	9	18

30

9. Evaluate $\sum_{i=1}^{6} 2Y$ given $\sum_{i=1}^{6} Y = 15$

B. False
1/2 point

10. $\Sigma(X+c)^2 = \Sigma X^2 + Nc^2$ (Where N is the number of elements summed)

A. True
B. False

11. B. False
1/2 point

11. $\Sigma 6(b+7U) = 6b+42\Sigma U$

A. True
B. False

78

12. Assume the convention where i, j, and k are the indices for the variables designated by the rows, columns, and observations within cells. For the data below, what is the numerical value of X_{321} ?

	1	2	3
1	59 47 58	58 38 18	74 49 23
2	91 24 80	31 42 12	93 36 48
3	89 47 62	78 36 88	62 69 79

Figure 10. A portion of a mastery test that is randomly parallel to that shown in figure 9.